International Journal of Occupational Safety and Ergonomics (*JOSE*) 2005, Vol. 11, No. 4, 399–407

Human Factors in Matching Images to Standards: Assimilation and Time Order Error

Achim Elfering

Department of Psychology, University of Berne, Berne, Switzerland

Objectives: *This study examines recognition performance to depend on image context and time order error. Recognition of standard images is a basic process in medical image analysis.*
Methods: *After the presentation of a standard square, 20 subjects identified the standard within a variety of 7 squares. The choice was between the standard and either 3 smaller and 3 larger squares, 5 smaller and 1 larger square, or 5 larger and 1 smaller square (context conditions).*
Results: *Multilevel regression analysis showed large individual differences in judgments (*$P < .001$*). Context induced assimilation of judgments to the medium-sized square within response options (*$P < .001$*). Negative time order error in rapid judgments caused an underestimation of the standard (*$P < .001$*).*
Conclusions: *Assimilation of judgments and time order error might be a threat to the reliability of medical image analysis. Some procedural recommendations are derived to reduce bias and increase patient safety in radiology.*

image reading context time order error patient safety

1. INTRODUCTION

The goal of this experiment was to determine time- and context-bound perceptual error in simple recognition of objects. Recognition of objects and structures is a basic process in perception of images, notably in analysis of medical images. The advanced aim of the study was to contribute to a better understanding of bias in medical image analysis. Diagnosing by means of medical image analysis is a very important part of radiologists' work. Medical treatment decisions often involve positron emission tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), or X-ray exams among other imaging options. After treatment, image analyses also help to monitor small but clinically important changes and contribute to overall outcome assessment. Reliable outcome assessment is essential for developing evidence-based medicine [1].

Therefore, reliable image analysis is a prerequisite in order to monitor quality of diagnoses and treatments, and reliable image analyses are closely connected to patient safety, especially in radiology [2]. Because developments in digital imaging will increase the relevance of medical image analyses within medicine, prevention of medical errors should concern reliability of image analysis to an even greater extent [3].

The importance of reliable image analysis is, for instance, in diagnosis and treatment of spinal disorders especially high. Within the complex structure of the spinal column, radiologists often have to judge and compare the size of highly similar tissues and structures according to categorical rating systems [4]. Thereby, reliability of radiologists' ratings of disk bulging/herniation and disk degeneration is often only moderate, and many factors threaten the reliability [5, 6]. Characteristics

The author is especially grateful to Prof. Dr. J. Hodler and Prof. Dr. N. Boos for their helpful comments on earlier versions of the manuscript.

Correspondence and requests for offprints should be sent to Achim Elfering, Department of Psychology, University of Berne, Muesmattstr. 45, CH-3000 Berne 9, Switzerland. E-mail:<achim.elfering@psy.unibe.ch>.

of the equipment, the categorical system, and human factors are potential biasing factors [7, 8, 9].

Human factors that threaten reliability of diagnoses are *cognitive errors* that include failures in perception (e.g., non-detection of a degenerated disk), failed heuristics (e.g., satisfaction of search error: the tendency to stop searching for distinctive disk-features after finding one), and biases (e.g., confirmation bias: the tendency to look for confirming evidence to support a diagnosis rather than look for disconfirming evidence to refute it, despite the latter often being more persuasive and definitive). Collectively, these have been referred to as *cognitive dispositions to respond* (CDRs) [10]. Recently, Croskerry [10] and Graber et al. [11] provided a comprehensive overview of CDR in medical decision making. CDR that refer basically to *perception* also include error-based misclassification of structures [12, 13], e.g., classification of a disk as normal instead of bulging in categorizing the extent of a lumbar disk herniation using a four-categorical grading system [14]. Basically, in this task prototypic or "standard" disks that stand for the degree of disk bulging are compared to the disks on the patient's medical image. In other words, standard disks are matched to current disks on medical images. Although this task is typical in the diagnosis of spinal disorders, little is known about cognitive error in this task. The task demands a reliable recognition of standards within a variety of images. Meanwhile, time-related factors (e.g., slow versus fast judgmental style) and contextual factors (e.g., similarity of adjacent tissues) may constrict recognition performance.

The present study investigates two sources of biases in recognition performance: the influence of (a) time order error (see section 1.1.), and (b) assimilation effects in dependency of the relative sizes of comparison stimuli that should be matched to the standard.

1.1. Time Order Error

Time order error (TOE) appears in judgments of two stimuli presented in a sequence. Two images of objectively the same size are rated differently as the image presented first is rated to be smaller than the second one [15]. With more than 500 ms between presentation of standard and comparison stimuli, and presentation of standard for more than a second, the standard is systematically underestimated with respect to size. The underestimation of the first stimulus is called negative TOE. In a recognition task, TOE should relate to processing time in recognition, that is the period of time and the duration of inspection of various comparison stimuli. TOE should be larger in rapid judgments, because longer response latencies are likely to involve more complex cognitive processes and therefore play a more prominent role than TOE per se. Hence, response latency should be inversely related to recognition performance, and rapid judgments should underestimate the standard.

In addition, it is a common finding that negative TOE increases with stimulus magnitude [15]. Thus, in this study, negative TOE should be larger in a larger standard than in a smaller one.

1.2. Assimilation Effects

The second factor of bias under study is context. In matching the standard with comparison cues, surrounding cues that are highly similar may systematically bias the matching decisions. Similar structures form a "context" that biases the recognition performance. There exists good evidence that recognition judgments are biased towards the medium magnitude of the stimulus variety. In other words, out of the variety of comparison stimuli that differ in size, individuals tend to choose a comparison stimulus of median size to match the standard—irrespective of the actual size of the standard. It has been shown before that individuals assimilate their recognition judgments to the midpoint of matching options; this assimilation effect has been consistently replicated in visual recognition of simple stimuli [16], and in recognition of rather complex stimuli like faces [17, 18].

Assimilation should therefore bias the matching of a standard when most optional images for matching are more extreme than the standard. In this study, the choice for matching was between the standard and either three smaller and three larger squares (contextual control condition), five

smaller and one larger square, or five larger and one smaller square (context conditions). In context conditions individuals were expected to choose a stimulus (to match the standard) that deviates towards the medium-sized value within options.

1.3. Research Questions

The broader aim of the study was to contribute to a better understanding of basic perceptual bias in medical image analysis. Simple recognition of standards is a basic process within categorization of tissue from medical images. TOE and stimulus context are supposed to systematically bias recognition performance. Three specific hypotheses were tested:

- TOE should be larger in rapid judgments compared to more delayed judgments;
- Negative TOE should be more expressed in larger compared to smaller standards; and
- Matching decisions should be biased towards the mean within options, i.e., assimilation in judgments is expected.

2. METHOD

2.1. The Experiment

As the cognitive effects under study are rather basic, the test of the hypotheses included a perception experiment including simple two-dimensional stimuli and students as participants. The intention was to avoid all individual and stimuli-bound factors that come into play in judgments of more medically meaningful complex stimuli (e.g., level of job experience and appropriateness of medical image resolution).

2.2. Subjects

Twenty right-handed undergraduate students (15 women and 5 men, mean age = 28.4 years, *SD* = 7.0) volunteered to participate. Except for prescription glasses or lenses (n = 9) no vision problems were known. All participants reported good visual acuity in viewing objects on the monitor screen. In previous studies, no gender effects on visual recognition performance were observed [16].

2.3. Apparatus

All stimuli were presented on a touch-screen device that also registered the responses. Using touch-screen interfaces, where users navigate a computer system by touching the screen, is the most simple, intuitive, and easiest way to interact and has the lowest potential to deflect attention with interface procedures or issues not associated with the task. The experiment was conducted with stimuli presented on a 15″ monitor with a touch-sensitive screen controlled by an IBM computer.

2.4. Stimuli

Fifteen red squares of different size were the stimuli. The sides of these squares varied (in 1-mm increments) from 12 to 26 mm. Trials included six different conditions with two standards and three contextual comparison sets (Table 1). The squares were projected on a grey background. Subjects sat 60 cm in front of the touch-screen monitor, so the resulting visual angles were between 1.15° and 2.48°. In previous studies, it was assured that differences between stimuli were above visual difference threshold [16].

TABLE 1. Standards and Comparison Stimuli of the Recognition Task

Stimuli (mm sides)	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Small standard square						■									
Context: control			■	■	■	■	■	■	■						
Smaller squares	■	■	■	■	■	■	■								
Larger squares					■	■	■	■	■	■	■				
Large standard square										■					
Context: control							■	■	■	■	■	■	■		
Smaller squares					■	■	■	■	■	■	■				
Larger squares									■	■	■	■	■	■	■

Notes. Stimuli varied in steps of 1-mm length of sides. Context control: comparison stimuli were arranged symmetrically around standard squares. Context conditions: with reference to the standards sizes, context test series included predominantly smaller comparison stimuli, or predominantly larger comparison stimuli.

2.5. Procedure

Subjects completed a 10-trial training session with acoustic feedback to become familiar with the task. During training, only three comparison stimuli were presented. After training subjects were told to expect an increase in task difficulty as seven comparison stimuli would be used and no feedback would be offered. Every trial started with a fixation cross presented in the center of the screen. Touching the cross started the trial and the standard was presented for 2000 ms before it disappeared and a blank screen appeared. After 1 s the comparison stimuli were presented in a cycle around the center of the screen (Figure 1). Subjects then chose a comparison stimulus to match the standard presented before. Choice and response latency were recorded as subjects touched a comparison stimulus at the screen. A choice could not be reversed. The sequence of the two standards in 126 trials and the monitor positions of the seven comparison stimuli were balanced. Trial intervals lasted 5 s, so the test session took about 30 min.

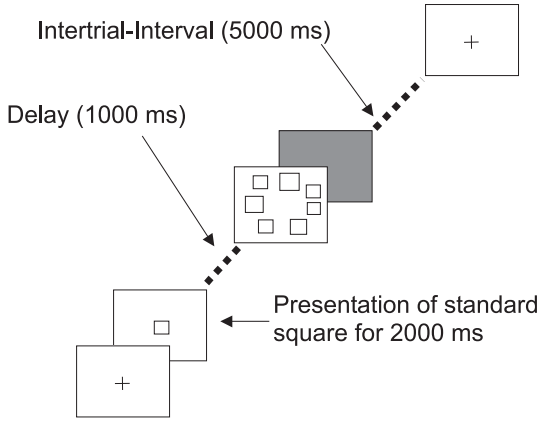

FIGURE 1. Matching of comparison stimuli to previously presented standard in delayed-matching-to-standard task (DMTS).

2.6. Data Analysis

Data contained information at the person- and the trial-level, with trials nested within persons. Previous studies demonstrated that subjects adopted different strategies when facing a change in context. Some subjects maintained previous standards, others switched rapidly and completely to the new context, and still others followed a compromise between these extremes [19, 20]. Individual variability in judgments causes statistical problems. On the one hand, a simple aggregation of trial-related information at the subject level would lead to loss of information and power. On the other hand, analyses on the trial-level (disaggregated data) would lead to an inflated

data set and spurious significances may result if all trials are treated as independent observations [21]. To deal with this problem, a multilevel linear regression model [22] approach was employed. It makes it possible to test the influence of trial-related variables and person-related measures, as well as cross-level interactions of trial- and person-related variables. The dependent variable in multilevel regression analysis was the side length of the square that was chosen to match the standard stimulus. As 20 subjects judged each of the six conditions (two standard stimuli with three context conditions each) 21 times, the total number of trials was 2520. The length of the matching square was regressed on standard stimuli, context condition, and response latency. The three two-way and the three-way interaction terms were also included into the full-factorial model. All predictor variables were centered to prevent multicollinearity between predictor variables and interaction terms [21]. The multilevel analyses were done with MLwiN software version 1.10 (Multilevel Models Project, London, UK). For all other analyses SPSS software package version 11.0 was used. A P value lower than .05 was considered significant.

3. RESULTS

Overall, participants tended to choose comparison stimuli physically smaller than the target stimuli. In other words, there was an overall tendency to underestimate the targets under all experimental conditions. Underestimation of the target stimuli is what one would expect from negative TOE. The 17-mm target stimulus was estimated to be 16.54 mm (SD = 1.32), and the 21-mm target stimulus was estimated to be 20.21 mm in length of sides (SD = 1.23). In line with this general tendency of underestimation, performance in terms of identification of the target was low. Twenty subjects in 2520 trials showed a hit rate of 26.5%. The hit rate was lower, if comparison stimuli were mainly smaller than the target (smaller context: 21.4%), and highest when comparison stimuli were mainly larger than the target (larger context: 30.7%), with the symmetrical contextual control condition lying in between (control context: 27.1%). Noteworthy, hit rate in comparably fast decisions, i.e., response latency below the median value of 2000 ms, was even lower (smaller context: 18.4%, control context: 22.4%, and larger context: 29.2%). Accordingly, performance data were comparably better in those trials with longer response latencies, i.e., in that half of trials with latencies longer than 2000 ms (smaller context: 24.2%, control context: 31.5%, and larger context: 32.3%). Taken together, Chi-square tests showed recognition performance to systematically depend on the stimulus context ($P < .001$) and the response latency ($P < .01$).

3.1. Multilevel Regression Analyses—Decomposition of Variance

Analyses started with the calculation of a variance components model in order to decompose the amounts of variance in the dependent variable explained by the situation level and the person level (estimation of the intra-class correlation, ICC). The ICC represents the proportion of the variance in the dependent variable explained by the person level [21]. The variance components model (with standard size as the only explanatory variable) yielded an ICC of .14 ($P < .001$). Thus, 14% of variation was on differences between individuals rather than situational factors, while 86% was at the trial-level.

3.2. Multilevel Regression Analyses—Fixed Effects

Table 2 shows the coefficients of the full regression model including all predictor variables and interactions. Not surprisingly, there was a strong effect of the standard size. The squares that were chosen to match the standard differed largely on whether the small or the large standard square was presented in a trial ($P < .001$). In accordance with the expectation of assimilation to the middle-sized stimuli of the context series, subjects chose a smaller square to match the standard when five comparison stimuli were actually smaller. Matched stimuli were larger in the symmetrical condition and they were largest when five comparison stimuli were actually larger than the standard ($P < .001$).

TABLE 2. Multilevel Model Prediction of the Size of the Square That Matched the Standard Square

Predictor Variables	Param.	*SE*
Fixed effects		
Intercept	18.380	0.098 [c]
Level 1		
Standard size	0.909	0.026 [c]
Context	0.274	0.027 [c]
Response latency	0.095	0.016 [c]
Level 1 Interactions		
Standard size x Context	–0.002	0.005
Standard size x Response latency	–0.011	0.008
Context x Response latency	0.030	0.008 [b]
Standard size x Context x Response latency	0.011	0.003 [c]
Random effects		
Variation intercept	0.184	0.061 [c]
Slope variation standard size	0.011	0.004 [b]
Slope variation context	0.012	0.004 [b]
COV standard size, intercept	0.005	0.011
COV context, intercept	–0.023	0.013
COV standard size, context	0.006	0.003 [b]
2log Likelihood (IGLS)	7373.89	

Notes. Sample size: $N = 2520$ from 20 subjects. Param.—fixed parameter estimates, IGLS—Iterative Generalised Least Squares. After the standard errors, the Wald-Test significance level (parameter estimates/standard error) is indicated with letters: a < .05, b < .01, c < .001, two-sided. Random effects—variance and covariance estimates of parameters that are allowed to vary on level 2. The Wald-Test is one-sided for variances (VAR) and two-sided for covariance estimates (COV). Codings for context: 0—symmetrical control condition (mid-stimulus matches standard size), –2—smaller comparison stimuli (mid-stimulus is 2 mm smaller than the standard), 2—larger comparison stimuli (mid-stimulus is 2 mm larger than the standard).

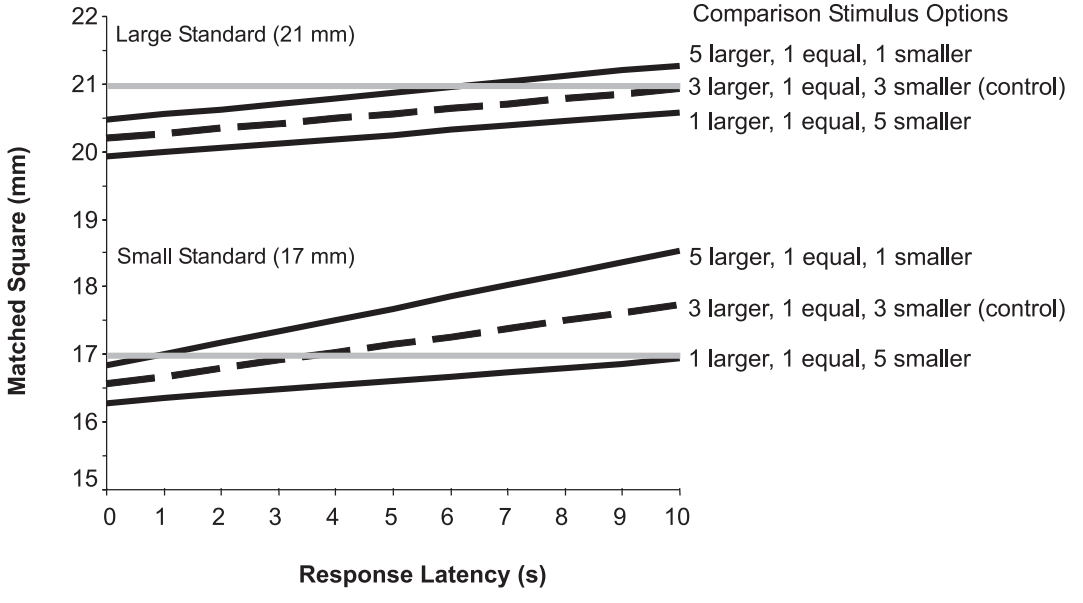

FIGURE 2. Regression lines as predicted from multilevel regression analysis. Stimulus choice (in mm side of squares) as a function of standard size, context, and response latency.

However, there was also a tendency to underestimate both standard squares that had 17- and 21-mm sides (see Figure 2). As expected from negative TOE, this tendency was inversely related to response latencies ($P < .001$). Context effects increased with response latency ($P < .001$); however, correspondence of context effect with response latency was higher in matching of the smaller standard than in matching of the larger standard ($P < .05$).

3.3. Multilevel Regression Analyses—Random Effects

Individuals differed in intercept, the slope of the standard size effect, and the slope of the assimilation effect. The significant estimate of variance in intercepts indicates that significant differences between individuals in mean judgments remain, even when all predictors entered the model. The size of the standard and context showed a significant random effect on slopes, i.e., slopes differing between individuals. Positive covariance ($P < .05$) between slopes of the standard size and context indicated those subjects who showed stronger assimilation (steeper slope of context predictor) to judge the standards to be more distinct in size (steeper slope of the standard size slope).

4. DISCUSSION

Diagnostic errors in radiology occur rather frequently. In the USA 30% of all medical malpractice lawsuits concern these errors [12]. A high percentage of these errors are perceptual misses, mostly subtle details in images that are not correctly detected and classified. Smith [23] estimated that 60% of diagnostic errors in radiology are perceptual.

To avoid perceptual errors, first the use of only high quality radiographs with adequate positioning should be standard. Second, radiologists should compare current radiographs with previous and standard radiographs in order to increase their sensitivity to subtle changes. In medical image analysis, most of the tasks include an inherent comparison of highly similar stimuli. Even though standard comparison scans are helpful, and radiologists are taught to compare [24], in practice, there exists only a moderate use of standards [25]. When using standards, errors may be located at three levels: scanning, recognition, and decision [26]. The goal of this experiment was to determine why errors were made in recognition. The results showed unique and joint effects of TOE and stimulus context on visual recognition performance. TOE was negative and restricted to a couple of seconds. Moreover, recognition performance was biased by assimilation towards the medium size of comparison stimuli. The direction and interdependencies of TOE and stimulus size agree with a recent model of TOE [12] that would expect larger TOE in larger standards. Individual variability was lower in TOE than in assimilation effects and, therefore, appeared to be a more basic perceptual process. Common recommendations to use heuristics for perceptual organization and search based on image characteristics, e.g., symmetry [27], should also consider TOE and assimilation to increase the reliability of medical image analysis. By now, picture archiving and communication systems (PACS) in the so-called dynamic soft-copy mode are standard and allow synchronous comparison of various images [28]. Today and in the near future, however, new technology will not only increase work flow but it will also increase work pace and workload because of an increasing demand [28], which makes time pressure and rapid judgments more likely. Rapid reading of medical images would increase the probability of TOE. Despite increasing time pressure, a major point from the effect of TOE shown in this study is that radiologists should not read images in a hurry.

Third, radiologists should repeat reading radiographs several times [12]. Note that radiologists tend to disagree with themselves as much as 20% of the time [29]. The results point to TOE and contextual cues as potential factors that contribute to intra-individual variability in performance. The results also show that individuals differ very much in recognition performance, and this supports the forth recommendation to increase patient safety in radiology. Whenever possible, medical images should be read by

several individuals simultaneously, and should be regularly discussed in training and education.

5. CONCLUSIONS

Radiologists should be aware of TOE and assimilation effects in categorization of similar structures. Whenever possible, medical images should be compared to standards. Assimilation of judgments should be controlled by repetition of readings in different contexts, most suitable in completely balanced trials. Medical images should be routinely read by multiple readers. In direct comparison of radiographs, e.g., baseline and follow-up radiographs, medical images should be judged several times with balanced order to level out negative TOE. Radiologists should avoid being pressed for time during judgment, as rapid judgments are more susceptible to TOE.

REFERENCES

1. Mannion AF, Elfering A. Predictors of surgical outcome. In: Boos N, Aebi M, editors. Spinal disorders—fundamentals of diagnosis and treatment. Berlin, Germany: Springer. In press.
2. Berlin L. Malpractice issues in radiology—defending the "missed" radiographic diagnosis. AJR 2001;176:317–22.
3. Espinosa J, Nolan T. Reducing errors made by emergency physicians in interpreting radiographs: longitudinal study. BMJ 2000;320:737–40.
4. Elfering A, Semmer NK, Birkhofer D, Zanetti M, Hodler J, Boos N. Risk factors for lumbar disc degeneration: a 5-year prospective MRI study in asymptomatic individuals. Spine 2002;27:125–34.
5. Parent E, Battié MC, Videman T. Quantitative MRI measures of disc bulging/herniation: development and reliability [paper presented at the 31st Annual ISSLS Meeting, Porto, Portugal]; 2004.
6. Raininko R, Manninen H, Battié MC, Gibbons LE, Gill K, Fisher LD. Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine. Spine 1995;20:1029–35.
7. Garland LH. Studies on the accuracy of diagnostic procedures. AJR 1959;82:25–38.
8. Carmody DP, Nodine CF, Kundel HL. An analysis of perceptual and cognitive factors in radiographic interpretation. Perception 1980;9:339–44.
9. Kundel HL. Perception errors in chest radiography. Seminars in Resp Med 1989;10:203–10.
10. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003,78:775–80.
11. Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: what's the goal? Acad Med 2002;77:981–92.
12. Berlin L. Malpractice issues in radiology: perceptual errors. AJR 1996;167:587–90.
13. Berlin L, Hendrix RW. Malpractice issues in radiology: Perceptual errors and negligence. AJR 1998;170:863–7.
14. Pfirrmann CWA, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 2001;26:1873–8.
15. Hellström A. Comparison is not just subtraction: effects of time- and space-order on subjective stimulus difference. P & P 2003;65:1161–77.
16. Elfering A. Psychophysikalische Methoden und Ergebnisse in der Bezugssystem-forschung: Die Rolle des Gedächtnisses im Reizgeneralisationsversuch [doctoral dissertation]. University of Frankfurt, Frankfurt, Germany; 1997.
17. Spetch MA, Cheng K, Clifford CWG. Peak shift but not range effects in recognition of faces. Learn Motiv 2004;35:221–41.
18. Webster MA, Kaping D, Mizokami Y, Duhamel P. Adaptation to natural facial categories. Nature 2004;428:557–61.
19. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Rad 1996;3:891–7.
20. Parducci A. Direction of shift in the judgment of single stimuli. J Exp Psychol 1956;51:169–78.
21. Hox JJ. Multilevel analysis. Mahwah, NJ, USA: Erlbaum; 2002.

22. Goldstein H, Browne W, Rabash J. Tutorial in biostatistics: multilevel modelling of medical data. Statist Med 2002;21:3291–315.
23. Smith MJ. Error and variation in diagnostic radiology. Springfield, IL, USA: Thomas; 1967.
24. American College of Radiology. Mammography quarterly standards act of 1992. Retrieved May 17, 2004, from://www.acr.org
25. Carmody DP, Kundel HL, Nodine CF. Comparison scans while reading chest images: taught but not practiced. Invest Radiol 1984;19:462–6.
26. Kundel HL, Nodine CF, Carmody DP. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Invest Radiol 1978;13:175–81.
27. Wackenheim A, Zollner G. Symetrie, asymetrie et dissymetrie. Ann Radiol 1987;30:60–4.
28. Reiner BI, Siegel EL, Siddiqui K. Evolution of the digital revolution: a radiologist perspective. Journal of Digital Imaging 2003;16:324–30.
29. Yerushalmy J. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. Radiol Clin North Am 1969;7:381–92.